Linking Provider Specialty and Outpatient Diagnoses in Medicare Claims Data: Data Quality Implications

Vojtech Huser[1] Nick D. Williams[1] Craig S. Mayer[1]

[1] Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, United States

Address for correspondence Vojtech Huser, MD, PhD, Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, United States (e-mail: vojtech.huser@nih.gov).

Appl Clin Inform 2021;12:729–736.

Abstract

Background With increasing use of real world data in observational health care research, data quality assessment of these data is equally gaining in importance. Electronic health record (EHR) or claims datasets can differ significantly in the spectrum of care covered by the data.

Objective In our study, we link provider specialty with diagnoses (encoded in International Classification of Diseases) with a motivation to characterize data completeness.

Methods We develop a set of measures that determine diagnostic span of a specialty (how many distinct diagnosis codes are generated by a specialty) and specialty span of a diagnosis (how many specialties diagnose a given condition). We also analyze ranked lists for both measures. As use case, we apply these measures to outpatient Medicare claims data from 2016 (3.5 billion diagnosis–specialty pairs). We analyze 82 distinct specialties present in Medicare claims (using Medicare list of specialties derived from level III Healthcare Provider Taxonomy Codes).

Results A typical specialty diagnoses on average 4,046 distinct diagnosis codes. It can range from 33 codes for medical toxicology to 25,475 codes for internal medicine. Specialties with large visit volume tend to have large diagnostic span. Median specialty span of a diagnosis code is 8 specialties with a range from 1 to 82 specialties. In total, 13.5% of all observed diagnoses are generated exclusively by a single specialty. Quantitative cumulative rankings reveal that some diagnosis codes can be dominated by few specialties. Using such diagnoses in cohort or outcome definitions may thus be vulnerable to incomplete specialty coverage of a given dataset.

Conclusion We propose specialty fingerprinting as a method to assess data completeness component of data quality. Datasets covering a full spectrum of care can be used to generate reference benchmark data that can quantify relative importance of a specialty in constructing diagnostic history elements of computable phenotype definitions.

Keywords

- data
- Medicare
- outpatient
- International Classification of Diseases, 10th Revision
- data quality

received
February 12, 2021
accepted after revision
June 22, 2021

DOI https://doi.org/10.1055/s-0041-1732404.
ISSN 1869-0327.

Background and Significance

In the last decade, the use of health care real world data (RWD) has been increasing when answering clinical research questions.[1–3] An ongoing research challenge is the assessment of data completeness or bias of RWD sources. Perfectly complete health history data over whole lifetime are rarely available. Instead, tools emerging in recent decades place emphasis on careful evaluation of fit for research use of a given database.[4] RWD can come from either administrative billing claims or from electronic health records (EHRs). While claim-based RWD databases typically cover the whole spectrum of care, EHR-sourced databases may reflect the spectrum of care covered by a given health care institution or larger delivery network of hospitals and clinics.[5–7] Examples of EHR-sourced databases include Stanford University's STARR (STAnford Research Repository) database,[8] or the Partners Healthcare Research Patient Data Registry.[9] We analyze diagnostic data in RWD by specialty. There is no existing study that would analyze specialty–diagnosis pairs across all diseases and specialties. A 2012 study by Wright et al[10] looked at EHR diagnostic data (problem list entries) and found that primary care providers added 82.3% of all problem list entries, despite writing only 40.4% of all EHR clinical notes.

Objectives

Our objective is to design a set of measures that characterize and quantify relationship between clinician's specialty and available diagnoses (to be used in research) in RWD. We used a comprehensive RWD database to characterize which diagnoses tend to be recorded in billing data by which medical specialties. When our set of measures is applied to a suitable RWD dataset, it can theoretically generate a reference benchmark data that would allow assessment of possibly missing diagnostic data in partial RWD databases. It is possible to quantitatively measure that a given specialty may be underrepresented in the dataset. For example, if dermatology practices are not part of an integrated delivery network, the corresponding network dataset will be missing dermatology visits and it will also be missing some dermatology-specific diagnoses made during such visits. We use event data of diagnosis and medical specialty pairs to characterize how care from a given specialty is recorded within the database. Such assessment can cast light on research suitability of a given dataset for a given research question.[11]

Methods

Our high level motivation for the study was to advance the informatics approaches to the assessment of data quality (DQ) and fit for use of RWD. We focus on assessing specialty–specific "data missingness" within diagnostic data. We envision applying this approach on various health care databases, preferably covering all age groups. To make our use case simpler, we only analyzed outpatient care within Medicare claims.

To clarify the context for our analysis, we define two types of datasets (full spectrum and partial spectrum datasets). We consider Medicare claims data to be a *full spectrum dataset* in terms of missing data with respect to a hypothetical complete record for insurance-covered health care events. A full spectrum dataset contains all care visits and their diagnostic data with no diagnostic data missing. Such an assumption can be made because the dataset originates from health insurance records. We acknowledge that, on higher level, loss (or change) of insurance and change of providers, not covered by that insurance, also has an impact. In contrast, we define a *partial spectrum dataset* as a database that covers only some visits that occur at care facilities that contribute data to the database (e.g., network of hospitals and clinics operated by some business entity). Such a definition implies that some visits and their diagnostic data may be missing from a partial spectrum dataset. For example, a research health care data warehouse maintained by a large academic medical center may be a partial spectrum dataset because patients may seek care at facilities that are not owned by the integrated delivery network and thus do not contribute data to the central research repository. By applying our set of characterization measures, it is theoretically possible to use a full spectrum dataset to generate benchmark reference data that could later be used for assessment of possibly missing data in a partial spectrum datasets.

Input Diagnostic Data

Our data consider clinical events paired with specialty of the provider making the diagnosis. We only considered a single specialty for each provider (sometimes referred to as primary specialty). As a use case, we used diagnostic data from claims of a large insurer in the United States. We used claims from the complete Medicare population (100% sample) available in the Virtual Research Data Center (VRDC) from Centers for Medicare and Medicaid Services (CMS).

We analyzed outpatient diagnostic events in Medicare claims for the calendar year 2016. Our initial pilot analyses used 1 month of data; using whole calendar years avoids seasonal variations in diagnoses (e.g., heat stroke occurring in summer and largely absent in winter). In the United States, outpatient professional billing claims link each diagnosis with a provider, and each provider who bills Medicare must declare a specialty.

Data extraction was done with SAS 9.4 (SAS Institute, Cary, North Carolina, United States) with additional analyses using R (Foundation for Statistical Computing, Vienna, Austria) on extracted data. We focus on characterization of typical specialty-specific diagnostic patterns and make no judgment on whether a given provider or specialty completely describes all diagnoses a patient may have. We fully acknowledge a health care billing context[12] and understand that only diagnoses somehow relevant to the currently billed procedure or service are recorded in the claim (i.e., comorbidities are not coded if they have no effect on the final bill).

Provider Specialties

There is no established mechanism for internationally harmonizing medical specialties, and minor differences by country may exist. A specialty can be considered in two contexts: self-declared specialty and exam-determined specialty.

In the United States, American Board of Medical Specialties is one example of terminology for exam-determined specialty. There are several terminologies for self-declared context. One is maintained by the American Medical Association for the purpose of conducting an annual survey. Others are used for billing or for government provider register of clinicians.

When applying for a National Provider Identifier from the National Plan and Provider Enumeration System, provider must select the Healthcare Provider Taxonomy Code (HPTC) that the provider determines most closely describes the health care provider's type/classification/specialization. Multiple codes can be declared but one of them must be declared as primary.

For billing, the HPTC code set is the only authorized standard that may be used in claim transactions to declare a specialty. It is maintained by National Uniform Claim Committee. The Code Set consists of two sections: individuals and groups of individuals, and nonindividuals. CMS maintains a crosswalk[13] that derives a CMS specialty code from HPTC level III code. Because CMS crosswalk included codes that are not specialties, we excluded some CMS specialty codes that indicated a facility rather than a medical specialty (e.g., "Portable X-ray supplier").

Analysis

After extracting paired diagnosis–specialty data from claims, we adopted two data perspectives: the first perspective used specialties as the main unit of analysis and the second perspective used diagnoses as the main unit of analyses.

Specialties

Prior research indicates that different specialties (primary care vs. specialists) do not equally contribute to maintaining a problem list.[10] We assumed a scenario that an imperfect problem list could possibly be built from claim diagnosis codes.[14] Given this assumption, we wanted to characterize specialties in terms of spectrum of diagnosis codes generated by them. In other words, how robust or vulnerable is the dataset to a missing specialty within a contributing integrated delivery network? In a partial spectrum dataset, if a given specialty is underrepresented in the data contributing network, what impact this may have on the captured diagnostic data? To partially characterize this phenomenon, we define *diagnostic span of specialty* as count of distinct diagnoses each specialty recorded. For example, diagnostic span of internal medicine may be very high compared with that of oral surgery. Diagnostic span also depends on the granularity of the underlying terminology for diagnoses. Because the diagnostic span of a specialty may partially depend on overall visit volume and overall diagnostic volume of a given specialty, we also characterize the proportion of all outpatient visits attributable to a specialty and the proportion of all diagnostic events generated by a specialty.

Diagnoses

In the second perspective that analyzed diagnoses, the motivation was to consider the following scenario: if a given specialty is missing (or is underrepresented) within the contributing delivery network, how much can its "missingness" be compensated for by the relatively complete presence of other specialties? Expressed yet differently, if a given disease diagnosis is not recorded due to missing data from specialty A, what other specialties (B, C, D, etc.) tend to diagnose that disease? We similarly created several measures to characterize diagnoses. We define *specialty span of a diagnosis* as count of distinct specialties that diagnose a given diagnostic code. For example, "Overweight (E66.3)" has large specialty span and is diagnosed by a large number of specialties compared with "Subacute and chronic vulvitis (N76.3)" that has a low specialty span. Next, to quantify the specialty span, we calculate separately for each diagnostic code the proportion of all diagnostic events attributable to a given specialty out of all diagnostic events for a given diagnostic code. This analysis allows us to distinguish diagnoses that are largely generated by a single or few specialties (top heavy), such as "Nonexudative age-related macular degeneration, bilateral, early dry stage (H35.3131)" (diagnosed 99% by ophthalmologists or optometrists), from top light diagnoses where volume from top five specialties is more spread out and not strongly dominated by a single specialty, such as "Nausea (R11.0)." For researchers replicating our analyses, both perspectives can be summarized as follows: for each outpatient diagnosis event generated by a provider, determine the specialty of that provider and assemble all "diagnosis code–specialty" pairs. Next, analyze both parts of each pair from both directions: number of distinct diagnosis codes per specialty and number of distinct specialties per diagnosis code. Finally, also quantify overall volumes (in absolute and relative terms) on visit (= date) level and event level for both parts of the pair, e.g., relative volume of diagnostic events for every specialty (see ►**Supplementary Appendix A** for an overview of all measures [available in the online version]).

Partial Spectrum Dataset Simulation

We chose simulation to study the effect of missing data by specialty. We compared the full outpatient dataset with five redacted datasets where we completely removed diagnostic events from a selected specialty (simulating a partial spectrum dataset). We compared full and redacted datasets using several measures: total number of distinct diagnoses within dataset, number of unique patient–diagnosis pairs, and number of diagnostic events.

Results

Provider Specialties

As of May 25, 2020, the CMS specialty list (as provided in the VRDC platform by CMS) contained 124 specialties, of which 98 were present in analyzed Medicare claims. The official publication of the CMS specialty crosswalk[13] (updated December 2020) uses HPTC codes effective as of April 2, 2018. From the VRDC-provided specialty list, we further excluded 14 CMS specialties that were present in data but that did not reflect a medical specialty (e.g., Centralized flu, or Mammography screening center). Supplemental file S1 provides the full list of CMS

Table 1 Overview of selected specialties showing diagnostic span, overall diagnostic volume, and visit volume (ordered by descending diagnostic span)

Specialty	Diagnostic span (rank)	Diagnostic volume (overall) %[a] (rank)	Visit volume % (rank)
Internal medicine	25,475 (1st)	16.004% (1st)	16.14% (1st)
Family practice	24,317 (2nd)	10.513% (2nd)	10.52% (2nd)
Nurse practitioner	21,508 (3rd)	6.094% (5th)	6.12% (6th)
Geriatric medicine	7,001 (19th)	0.323% (44th)	0.32% (44th)
Ophthalmology	5,859 (26th)	3.395% (8th)	3.33% (8th)
Radiation oncology	3,359 (48th)	0.485% (37th)	0.48% (40th)
Sleep medicine	1,369 (66th)	0.035% (65th)	0.03% (65th)
Geriatric psychiatry	649 (77th)	0.016% (71st)	0.01% (71st)
Psychologist	145 (80th)	0.006% (79th)	0.006 (79th)

[a]Diagnostic volume (overall) % is calculated as count of all diagnostic events for a specialty divided by count of all analyzed diagnostic events.

specialties with flags for "present in data" and "excluded from analysis" (supplemental files are available at https://github.com/lhncbc/CRI/tree/master/VRDC/project/specialty). The final analyzed list considered 82 specialties.

Input Diagnostic Data

There are two levels to consider: distinct diagnoses in terminology and actual patient level diagnostic events. First, in terms of International Classification of Diseases, 10th Revision Clinical Modification (ICD10-CM) terminology, the 2020 ICD10-CM terminology contains 95,958 active diagnosis codes (and 1,566 inactive codes). Because a given dataset may not utilize all available codes, the true starting point was a total of 35,233 distinct codes that were present in analyzed Medicare claims. Within this present-in data universe of codes, a further 384 codes were removed because they were only paired with excluded CMS specialties. The final analyzed set thus contained 34,794 ICD10-CM codes. Supplemental file S2 provides the full list of ICD-10 CM codes with flags for "present in data" and "paired with excluded CMS specialty code."

Second, in terms of patient level diagnostic event volume, we analyzed a total of 3,412,857,167 noninstitutional outpatient diagnostic events. This represents 74.5% of all 2016 diagnoses. The remaining diagnoses (not analyzed) were from the following claim types: institutional outpatient (17.13%), inpatient (4.39%), skilled nursing facility (1.45%), home health (1.29%), and hospice (0.8%). From this set of events (3.412 billion), a further 11.3% events were removed since they were paired with excluded CMS specialty.

Analysis of Specialties

Considering specialty as the unit of analysis, ►Table 1 shows diagnostic span and diagnostic volume for a small set of selected specialties. Supplemental file S-T1 contains data for all 82 analyzed specialties. The diagnostic span of a specialty ranges from 33 distinct diagnoses (medical toxicology) to 25,475 diagnoses (internal medicine). ►Table 1 showcases specialties with very wide diagnostic span (internal medicine, family practice, and nurse practitioner). Those specialties are also at the same time responsible for the largest proportion of overall diagnostic events. For example, family practice accounts for 10.5% of all outpatient diagnoses. The table also shows specialties with narrow span, such as psychologist (145 diagnoses) or sleep medicine (1,369 diagnoses). The median diagnostic span is 4,046 distinct diagnoses and the interquartile range is 1,889 to 6,611 distinct diagnoses.

We next looked at frequency of a diagnosis within a specialty. ►Table 2 shows the top five diagnoses for three selected specialties. Supplemental file S-T2 provides these data for all 82 specialties. The specialty headers in ►Table 2 repeat data on diagnostic span to demonstrate that we purposefully selected specialties with a different diagnostic span. ►Table 2 shows that specialties differ in distribution of diagnostic volume. For example, the cumulative share of top five diagnoses is high for sleep medicine (47.52%; can be considered top heavy) but low for pathology (7.52%). Further highlighting individual diagnoses, "pathology" has a large diagnostic span of 10,944 distinct diagnoses and the most frequent diagnosis (Unspecified chronic gastritis without bleeding [K29.50]) was only responsible for 1.67% of specialty's diagnostic volume. The diagnostic pattern (or scenario) for "pathology" is thus a relatively wide diagnostic span and at the same time not "top heavy." On the other hand, "sleep medicine" has nine times smaller diagnostic span of 1,369 distinct diagnoses and the most frequent diagnoses (Obstructive sleep apnea [G47.33]) accounts for 36.68% of specialty's diagnostic volume. It represents a different scenario of relatively narrow specialty diagnostic span and much more top heavy. Somewhat in between these scenarios is the third featured specialty of nephrology.

Analysis of Diagnoses

The second perspective used diagnosis as the unit of analysis. ►Table 3 shows specialty span and overall diagnostic volume for seven selected diagnoses. Supplemental file S-T3 contains the same data for all 34,794 analyzed ICD10-CM diagnoses. Specialty span of a diagnosis ranges between 82 (hypertension; I10) and 1. On the low end of the spectrum, a

Table 2 Top five diagnoses by diagnostic volume percent (within specialty)

Specialty	Diagnosis	Diagnostic volume (within specialty) (%)[a]	Cumulative volume (within specialty) (%)
Sleep medicine (span: 1,369, code volume %: 0.035%)	Obstructive sleep apnea (G47.33)	36.68%	36.67%
	Essential (primary) hypertension (I10)	4.29%	40.97%
	Snoring (R06.83)	2.24%	43.21%
	Sleep apnea, unspecified (G47.30)	2.20%	45.42%
	Hypersomnia, unspecified (G47.10)	2.11%	47.53%
Nephrology (span: 6,846, code volume %: 1.86%)	End-stage renal disease (N18.6)	16.98%	16.97%
	Chronic kidney disease, stage 3 (moderate) (N18.3)	7.08%	24.05%
	Acute kidney failure, unspecified (N17.9)	6.32%	30.37%
	Essential (primary) hypertension (I10)	5.59%	35.96%
	Dependence on renal dialysis (Z99.2)	4.97%	40.93%
Pathology (span: 10,944, code volume%: 1.20%)	Unspecified chronic gastritis without bleeding (K29.50)	1.67%	1.67%
	Essential (primary) hypertension (I10)	1.59%	3.26%
	Actinic keratosis (L57.0)	1.54%	4.79%
	Polyp of colon (K63.5)	1.36%	6.15%
	Benign neoplasm of ascending colon (D12.2)	1.36%	7.51%

[a]Diagnostic volume (within specialty) % is calculated as count of diagnostic events of a given diagnosis divided by count of all diagnostic events made by a given specialty.

Table 3 Specialty span and overall diagnostic volume of selected diagnoses

Diagnosis	Specialty span (rank)	Diagnostic volume (overall) % (rank)[a]
Hyperlipidemia, unspecified (E78.5)	80 (2nd)	1.106978% (5th)
End-stage renal disease (N18.6)	79 (3rd)	0.583976% (13th)
Muscle weakness (generalized) (M62.81)	77 (5th)	0.409923% (29th)
Congenital dilation of aorta (Q25.44)	8 (75th)	0.000034% (17,592nd)
Charcot joint, multiple sites (M14.69)	10 (73rd)	0.000033% (17,752nd)
Hypoplastic left heart syndrome (Q23.4)	7 (76th)	0.000033% (17,775th)
Typhoid meningitis (A01.01)	1 (82nd)	0.000001% (32,023rd)

[a]Diagnostic volume (overall) % is calculated as count of diagnostic events of a given diagnosis divided by the total count of all diagnostic events.

total of 4,762 (13.51% out of 35,233 analyzed diagnoses) are diagnosed by a single specialty. The median specialty span is 8 and interquartile range is 3 to 17.

The diagnoses shown in ►Table 3 represent a convenience sample that was selected to demonstrate the variability. It shows example diagnoses that are diagnosed by a large number of specialties (e.g., "Hyperlipidemia, unspecified" [E78.5; 80 specialties] or "End stage renal disease" [N18.6; 79 specialties]) as well as example diagnoses diagnosed by a small number of specialties (e.g., "Hypoplastic left heart syndrome" [Q23.4]: seven specialties, or "Typhoid meningitis" [A01.01]: one specialty). The specialty span of a diagnosis is naturally affected by the prevalence of the condition as well as intensity of care for a given condition (number of repeated visits and to what range of specialists). See the last column in ►Table 3 that shows the overall diagnostic volume which reflects both prevalence and repeated visits. Specialty span (column 2) is correlated with the overall diagnostic column (column 3) using Pearson's product–moment correlation method ($r^2 = 0.92$; correlation coefficient: 0.273 [confidence interval: 0.263–0.282], p-value of $<2.2e-16$). The larger is the diagnostic volume, the larger is the specialty span (i.e., more specialties diagnose it).

Looking at specialty spans across all diagnoses, we can say that 14,011 diagnoses (40.26% of all 34,794 diagnoses) are diagnosed by five or less specialties, and 79.53% of diagnoses are diagnosed by 20 or less specialties. For later analysis, we can loosely define a concept of *highly specialty-specific diagnosis*, where 40% or more of all diagnostic events are provided by a single specialty. Using this definition, 19,514 diagnoses (56.08%) are highly specialty-specific (e.g., Manic

Table 4 Top five specialties diagnosing a given diagnosis

Diagnosis	Specialty	Diagnostic volume (within diagnosis) (%)[a]	Cumulative volume (%)
Urinary tract infection, site not specified (N39.0) (span: 77, Dx volume %: 0.5575%)	Internal medicine	27.81%	27.81%
	Family practice	15.51%	43.32%
	Urology	10.99%	54.31%
	Nurse practitioner	10.85%	65.16%
	Emergency medicine	9.82%	74.98%
Rupture of artery (I77.2) (span:27, Dx volume %: 0.00020%)	Internal medicine	12.23%	12.23%
	Vascular surgery	11.39%	23.62%
	Diagnostic radiology	9.27%	32.89%
	Nurse practitioner	7.74%	40.63%
	General surgery	7.48%	48.11%
Manic episode in partial remission (F30.3) (span: 7, Dx volume %: 0.00011%)	Psychiatry	41.23%	41.23%
	Licensed clinical social worker	26.20%	67.43%
	Clinical psychologist	14.06%	81.49%
	Nurse practitioner	9.22%	90.71%
	Family practice	4.12%	94.83%

[a]Diagnostic volume (within a diagnosis) % is calculated as proportion of diagnostic events for a given diagnosis made by a given specialty (i.e., numerator is the count of diagnostic events of a given diagnosis made by a given specialty divided by the count of all diagnostic events of a given diagnosis).

episode in partial remission; F30.3). Obviously, all specialties diagnosed by a single specialty (shown earlier to be 13.5% of all diagnoses analyzed) are naturally also highly specialty-specific. On the opposite side, we can similarly define a concept of *multispecialty diagnosis*, where top five specialties generate less than 20% of volume of events for that diagnosis. With this definition, 1,595 diagnoses (4.58%) would be considered multispecialty. The thresholds we chose in such definitions are not based on any formal methodology and can be easily changed and results recomputed.

Similarly to the first perspective, we next looked at patterns of specialty frequencies (within a diagnosis). ►Table 4 presents a view of top specialties within a given diagnosis for three selected diagnoses. Full data for all diagnoses are available in Supplemental file S-T4. ►Table 4 repeats data on specialty span and shows diagnoses with wide as well as narrow specialty span. An example of diagnosis with a narrow specialty span and relatively top heavy by diagnosing specialty is "Manic episode in partial remission" (F30.3). Psychiatry is responsible for 41.23% of events with this diagnosis and the top three specialties making this diagnosis account for 81.49% of such diagnostic events (top heavy indicator). The other two examples in ►Table 4 show diagnoses that are not top-heavy in terms of diagnosing specialty. One has a wide diagnostic span and is relatively frequent (Urinary tract infection, site not specified; N39.0) and the other has a medium diagnostic span and is less frequent (Rupture of artery; I77.2). The point of ►Table 4's perspective is to show that specialty composition of a given diagnosis can vary significantly.

Reference information for each diagnosis (in Supplemental file S-T4) that is generated from full spectrum datasets can be useful to researchers in creation of diagnosis-driven phenotypes that are executed on partial spectrum datasets.[15] For example, absence of emergency medicine specialty data would reduce the volume of "Urinary tract infection, site not specified" diagnoses by 9.82% (see ►Table 4).

Partial Spectrum Dataset Simulation

In ►Supplementary Appendix A (available in the online version), we present results of simulation of partial-spectrum datasets. This additional analysis of comments on two tables is presented in ►Supplementary Appendix B. ►Supplementary Table S1 (available in the online version) shows on three different levels (population level, person level, and event level) drop in data completeness relative to the full, unreacted dataset. ►Supplementary Table S2 (available in the online version) demonstrates the difference on selected individual diagnoses.

Discussion

Our analysis is the first to analyze diagnosis–specialty pairs in claims data across all diseases. We provide descriptive data that show that some diagnoses are made by a single or few specialties. Our measures quantify an assertion that datasets that miss a specialty can provide a distorted diagnostic picture of a population. We demonstrate that diagnoses can range from multispecialty (somewhat resistant to such distortion) to highly specialty-specific (more prone to such distortion). Analysis of diagnosis–specialty pairs can offer new approaches to assessment of data completeness. The motivation for our work was to show that full or partial spectrum datasets (in terms of represented specialties) can differ and what measures can be used to quantify such differences. Our set of measures

based on diagnosis–specialty pairs advances the field of data assessment. Clear characterization of RWD is important when researchers need to pick the most suitable RWD database for a given research question.

When the approach is applied to a suitable reference dataset, it is possible to generate a look-up table for diagnoses with useful specialty information. Data on all diagnoses (available in Supplemental file S-T4; the most valuable output of our study) allow researchers to see which specialties are most relevant for any given diagnosis. Researchers could be asking the following questions: How complete is my dataset? On visit level, what visits may be missing by design (due to dataset origin)? On provider level, providers of which specialty may be partially missing from my dataset and to what extent? How important is complete picture of patient's comorbidities for my research analysis?

In addition to reference benchmark data that focus on individual diagnoses (see Supplemental file S-T4), it is also possible to characterize dataset's partial spectrum specialty coverage by directly looking at frequency distribution of provider specialties within the dataset: a *specialty snapshot of the dataset*. In Kahn et al's DQ framework,[16] which consists of conformance, completeness, and plausibility, such DQ check would fall under a completeness category (Are data values present?) and validation context (Is there agreement with an external reference?). Distribution of specialties should conform to valid external reference. Such checking can also be implemented in a rule-based DQ framework.[17] Sentinel network administrators also use the term database fingerprinting and the term in their framework would thus be *specialty fingerprinting*.[18,19] On implementation level, the specialty snapshot can be done on provider level or visit level (total number of distinct dermatologists in the dataset or number of visits to any dermatologist). Because not all providers are working full time, the visit-level snapshot is more accurate. We have added this feature as a new query to the Observational Health Data Science and Informatics DQ tool called Achilles. We consider incorporation of this additional DQ characterization into an established tool a significant contribution of our study to clinical research informatics. To develop this idea even further, in addition to absolute values' comparison of provider visit frequencies, a relative measure that uses a ratio of two specialties is also possible. For example, for every 1,000 visits to "family practice" specialty, we expect to observe 19.3 visits to "neurosurgery" specialty. We did not attempt to produce reference data for visit-level specialty snapshot and considered it out of scope of the current study mainly because the Medicare data reflect largely care for patients aged 65+. The goal of our work was to develop a new methodology for working with diagnosis–specialty event pairs on informatics level. As future work, we plan to apply this set of measures on multiple datasets, compare them to each other, and experiment with multiple reference benchmark data for specialty snapshot assessment.

Limitations

Our study has several limitations. First, we only considered outpatient diagnoses. Second, our input claims data come from Medicare and the benchmark results are only valid for senior population (age > 65). However, our emphasis is on developing a methodology rather than on producing the perfect reference benchmark data.

Third, our assumption that claims data represent a full spectrum dataset is a compromise. We fully acknowledge that insurer-based view of patient diagnoses has limitations (e.g., no data on self-pay care or data on health issues that do not lead to a visit).

Fourth, we only analyzed a single year time window (mostly because of query speed). Related to that is a more complex view (e.g., 5 years for chronic diagnoses) of the "missingness" versus "lateness." An inference about a chronic diagnosis can be wrong in two aspects: (1) present/absent binary axis or (2) inaccurate inferred date of onset (later than the accurate date). This complex temporal aspect was outside the scope of the presented study.

Finally, the diagnostic span of specialty and other analyses were conducted at the default, leaf level of granularity of the ICD10CM terminology. In the presented results, we did not attempt to group diagnoses into a higher level of aggregation. Because the purpose of the ICD10-CM coding is primarily billing, the level of granularity may not be uniform across specialties. Billing purpose may require granularity and distinction that is not important to observational research. Large diagnostic span of a specialty may simply be due to differing levels of ICD10-CM billing granularity by disease. There are numerous disease groupings developed over time and by several research teams. Such aggregations reduce the number of diagnostic codes because the research question is better answered on granularity optimized for research (not billing). To partially demonstrate possible grouping of ICD10-CM diagnostic codes, we piloted (just for the diagnostic span determination) an aggregation that used ICD10-CM mappings into SNOMED CT from the vocabulary layer of the Observational Medical Outcomes Partnership model. For example, "Motion sickness, initial encounter" (T75.3XXA) and "Motion sickness, subsequent encounter" (T75.3XXD) are both mapped to the same SNOMED CT concept of "Motion sickness" (SCTID: 37031009). Using this grouping, the span of 34,794 distinct ICD10-CM diagnoses in Medicare outpatient data are reduced to 11,328 distinct diagnostic groups. For the specialty of "internal medicine" the diagnostic span reduces from 25,475 ICD10-CM diagnoses to 9,914 grouped diagnoses.

Conclusion

DQ assessment of RWD is an evolving field that is constantly looking for new data evaluation perspectives. Our work pioneers one such new perspective that uses clinician's specialty. Analysis of diagnosis–specialty pair events reveals differences among specialties and specialty composition of individual diagnoses. We propose specialty fingerprinting as a method to assess data completeness. Datasets covering full spectrum of care can be used to generate reference benchmark data that can quantify relative importance of a specialty in constructing diagnostic history elements of computable

phenotype definitions. Our set of measures can be incorporated into existing DQ assessment tools. The results also help researchers better realize operating characteristics of health care billing data in research context.

Clinical Relevance Statement

A problem list component of an EHR system can facilitate advanced patient management. This article assumes possibility of building a problem list from claims data. Our results quantify contribution of different specialties to such a problem list.

Multiple Choice Questions

1. In U.S. claims, as of 2021, the following terminology is used to encode diagnoses:
   a. ICD10.
   b. ICD10CN.
   c. ICD10CM.
   d. ICD9CM.

   Correct Answer: The correct answer is option c. Since October 1, 2015, ICD10CM (modified version of ICD10 [international version]) is used in the United States. See https://www.cms.gov/Medicare/Coding/ICD10.

2. Specialty for a physician in U.S. National Provider Index is
   a. Determined by board certification exam.
   b. Self-declared by physician.
   c. Assigned by CMS.
   d. Based on annual survey conducted by American Medical Association.

   Correct Answer: The correct answer is option b. Each provider can select a single specialty that most closely matches their specialization.

Protection of Human and Animal Subjects

This study was declared not human subject research by the Office of Human Research Protection at National Institutes of Health.

Funding

None.

Conflict of Interest

None declared.

Acknowledgment

We would like to thank Dr. Laritza Rodriguez for providing comments on an earlier version of this manuscript. This research was performed by staff of the National Library of Medicine (NLM), National Institutes of Health (NIH), with support from NLM. The findings and conclusions in this article are those of the authors and do not necessarily represent the official position of NLM, NIH, or the Department of Health and Human Services.

References

1 Briere J-B, Bowrin K, Taieb V, Millier A, Toumi M, Coleman C. Meta-analyses using real-world data to generate clinical and epidemiological evidence: a systematic literature review of existing recommendations. Curr Med Res Opin 2018;34(12): 2125–2130
2 Bowrin K, Briere J-B, Levy P, Millier A, Clay E, Toumi M. Cost-effectiveness analyses using real-world data: an overview of the literature. J Med Econ 2019;22(06):545–553
3 Ramamoorthy A, Huang S-M. What does it take to transform real-world data into real-world evidence? Clin Pharmacol Ther 2019; 106(01):10–18
4 Food and Drug Administration. Real-world evidence. Accessed May 3, 2021 at: https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence
5 Edlinger D, Sauter SK, Rinner C, et al. JADE: a tool for medical researchers to explore adverse drug events using health claims data. Appl Clin Inform 2014;5(03):621–629
6 Cusick MM, Sholle ET, Davila MA, Kabariti J, Cole CL, Campion TR Jr. A method to improve availability and quality of patient race data in an electronic health record system. Appl Clin Inform 2020; 11(05):785–791
7 Joukes E, de Keizer NF, de Bruijne MC, Abu-Hanna A, Cornet R. Impact of electronic versus paper-based recording before EHR implementation on health care professionals' perceptions of EHR use, data quality, and data reuse. Appl Clin Inform 2019;10(02): 199–209
8 STAnford Research Repository (STARR) Tools Accessed Feb 15, 2021 at: https://med.stanford.edu/starr-tools.html
9 Murphy SN, Gainer V, Chueh HC. A visual interface designed for novice users to find research patient cohorts in a large biomedical database. AMIA Annu Symp Proc 2003;2003:489–493
10 Wright A, Feblowitz J, Maloney FL, Henkin S, Bates DW. Use of an electronic problem list by primary care providers and specialists. J Gen Intern Med 2012;27(08):968–973
11 Sundaresan AS, Schneider G, Reynolds J, Kirchner HL. Identifying asthma exacerbation-related emergency department visit using electronic medical record and claims data. Appl Clin Inform 2018; 9(03):528–540
12 Mayer C, Williams N, Huser V; National Institutes of Health. Loss of diagnostic data due to claim form limitations. Poster presented at AMIA 2021 Virtual Informatics Summit; March 24, 2021
13 Data.CMS.gov. Crosswalk Medicare provider/supplier to HPTC. Accessed April 21, 2021 at: https://data.cms.gov/Medicare-Enrollment/CROSSWALK-MEDICARE-PROVIDER-SUPPLIER-to-HEALTHCAR-E/j75i-rw8y/
14 Wright A, Pang J, Feblowitz JC, et al. A method and knowledge base for automated inference of patient problems from structured data in an electronic medical record. J Am Med Inform Assoc 2011;18 (06):859–867
15 Cameron CB. A user's guide to computable phenotypes. Accessed April 10, 2021 at: https://dcricollab.dcri.duke.edu/sites/NIHKR/KR/Blake_Users_Guide_to_Computable_Phenotypes.pdf
16 Kahn MG, Callahan TJ, Barnard J, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. EGEMS (Wash DC) 2016;4(01):1244
17 Wang Z, Talburt JR, Wu N, Dagtas S, Zozus MN. A rule-based data quality assessment system for electronic health record data. Appl Clin Inform 2020;11(04):622–634
18 Sentinel. Standardization and querying of data quality metrics and characteristics for electronic health data. Accessed Feb 14, 2021 at: https://www.sentinelinitiative.org/sites/default/files/Methods/Standardization_and_Querying_of_Data_Quality_Metrics.pdf
19 Huser V, Kahn MG, Brown JS, Gouripeddi R. Methods for examining data quality in healthcare integrated data repositories. Pac Symp Biocomput 2018;23:628–633

